INTERNATIONAL HPC ACTIVITIES

Jack Dongarra

University of Tennessee

Oak Ridge National Laboratory

State of Supercomputing Today

- **Pflops (> 10^{15} Flop/s) computing fully established with 95 systems.**
- **Three technology architecture possibilities or "swim lanes" are thriving.**
  - Commodity (e.g. Intel)
  - Commodity + accelerator (e.g. GPUs) (93 systems)
  - Lightweight cores (e.g. ShenWei, ARM, Intel's Knights Landing)
- **Interest in supercomputing is now worldwide, and growing in many new markets** (around 50% of Top500 computers are used in industry).
- **Exascale (10^{18} Flop/s) projects exist in many countries and regions.**
- **Intel processors have largest share, 91% followed by AMD, 3%.**
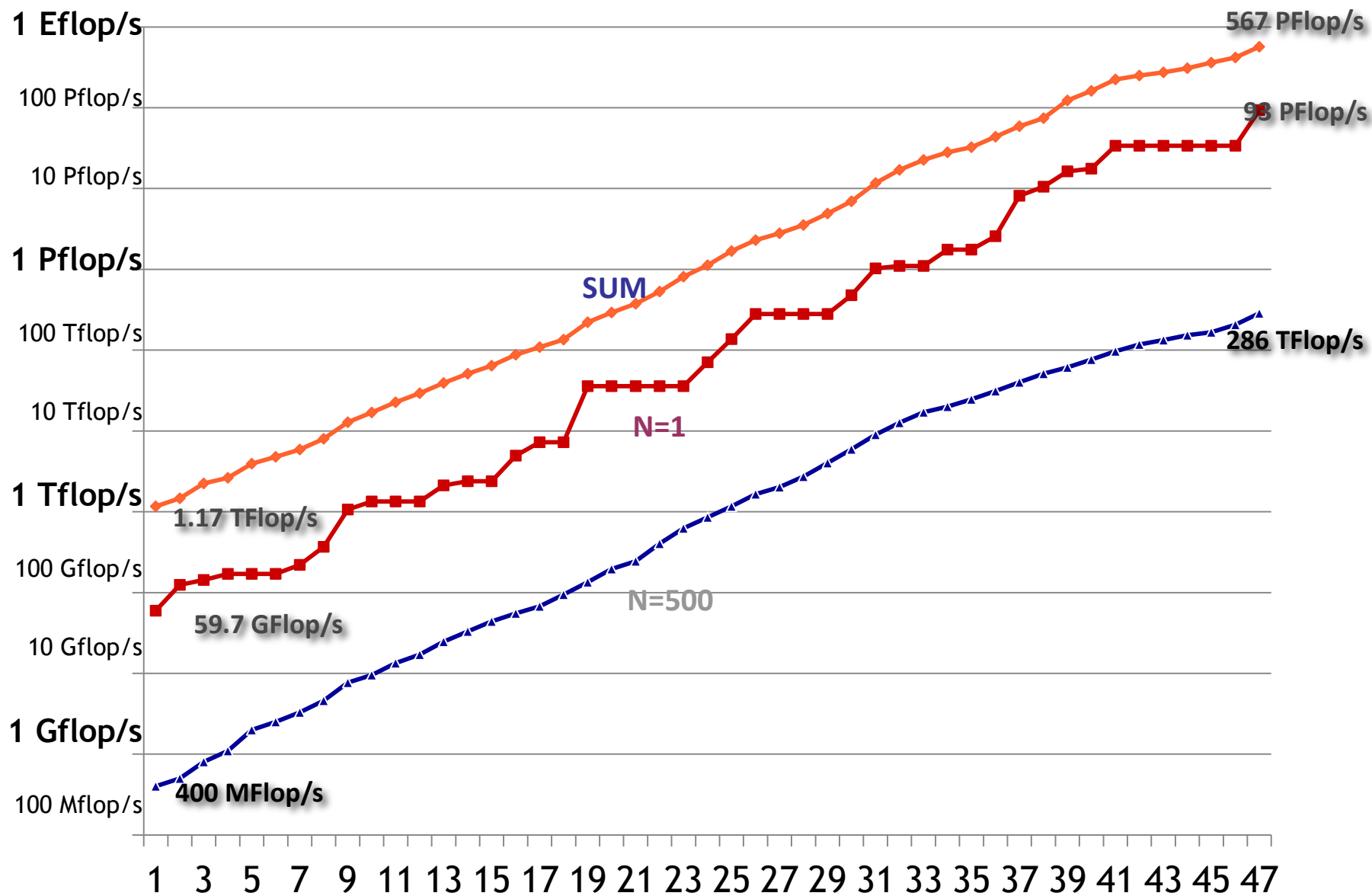
H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
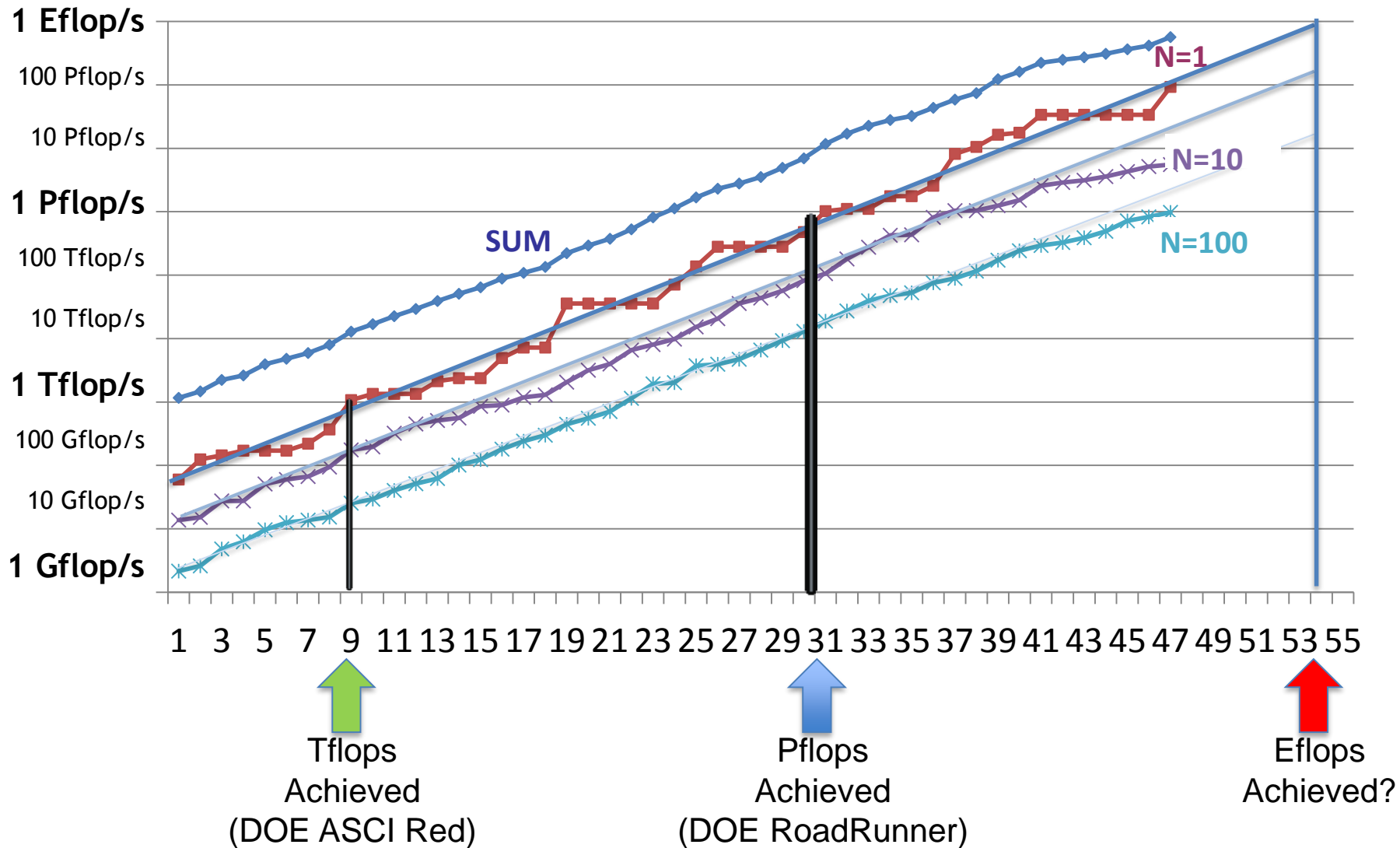- Yardstick: Rmax from LINPACK MPP

$$Ax=b, \text{ } \textit{dense problem}$$
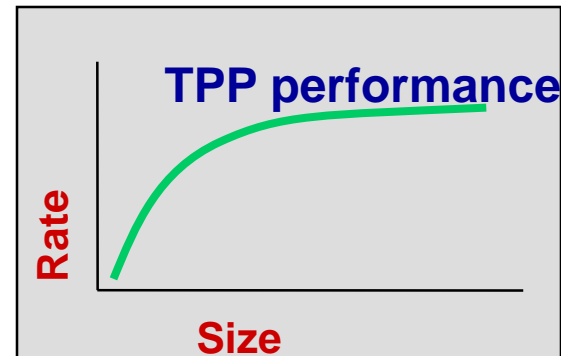
- Updated twice a year
  SC'xy in the States in November
  Meeting in Germany in June

- All data available from **www.top500.org**

3

Performance Development of HPC over the Last 24 Years from the Top500

567 PFlop/s

93 PFlop/s

286 TFlop/s

SUM

N=1

N=500

1.17 TFlop/s

59.7 GFlop/s

400 MFlop/s

PERFORMANCE DEVELOPMENT

June 2016: The TOP 10 Systems

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C) + Custom	China	10,649,000	93.0	74	15.4	6.04
2	National Super Computer Center in Guangzhou	Tianhe-2 NUDT, Xeon (12C) + IntelXeon Phi (57c) + Custom	China	3,120,000	33.9	62	17.8	1.91
3	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7, AMD (16C) + Nvidia Kepler GPU (14c) + Custom	USA	560,640	17.6	65	8.21	2.14
4	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16C) + custom	USA	1,572,864	17.2	85	7.89	2.18
5	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8C) + Custom	Japan	705,024	10.5	93	12.7	.827
6	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16C) + Custom	USA	786,432	8.16	85	3.95	2.07
7	DOE / NNSA / Los Alamos & Sandia	Trinity, Cray XC40,Xeon (16C) + Custom	USA	301,056	8.10	80	4.23	1.92
8	Swiss CSCS	Piz Daint, Cray XC30, Xeon (8C) + Nvidia Kepler (14c) + Custom	Swiss	115,984	6.27	81	2.33	2.69
9	HLRS Stuttgart	Hazel Hen, Cray XC40, Xeon (12C) + Custom	Germany	185,088	5.64	76	3.62	1.56
10	KAUST	Shaheen II, Cray XC40, Xeon (16C) + Custom	Saudi Arabia	196,608	5.54	77	2.83	1.96
500	Internet company	Inspur Intel (8C) + Nnvidia	China	5440	.286	71		

Countries Share
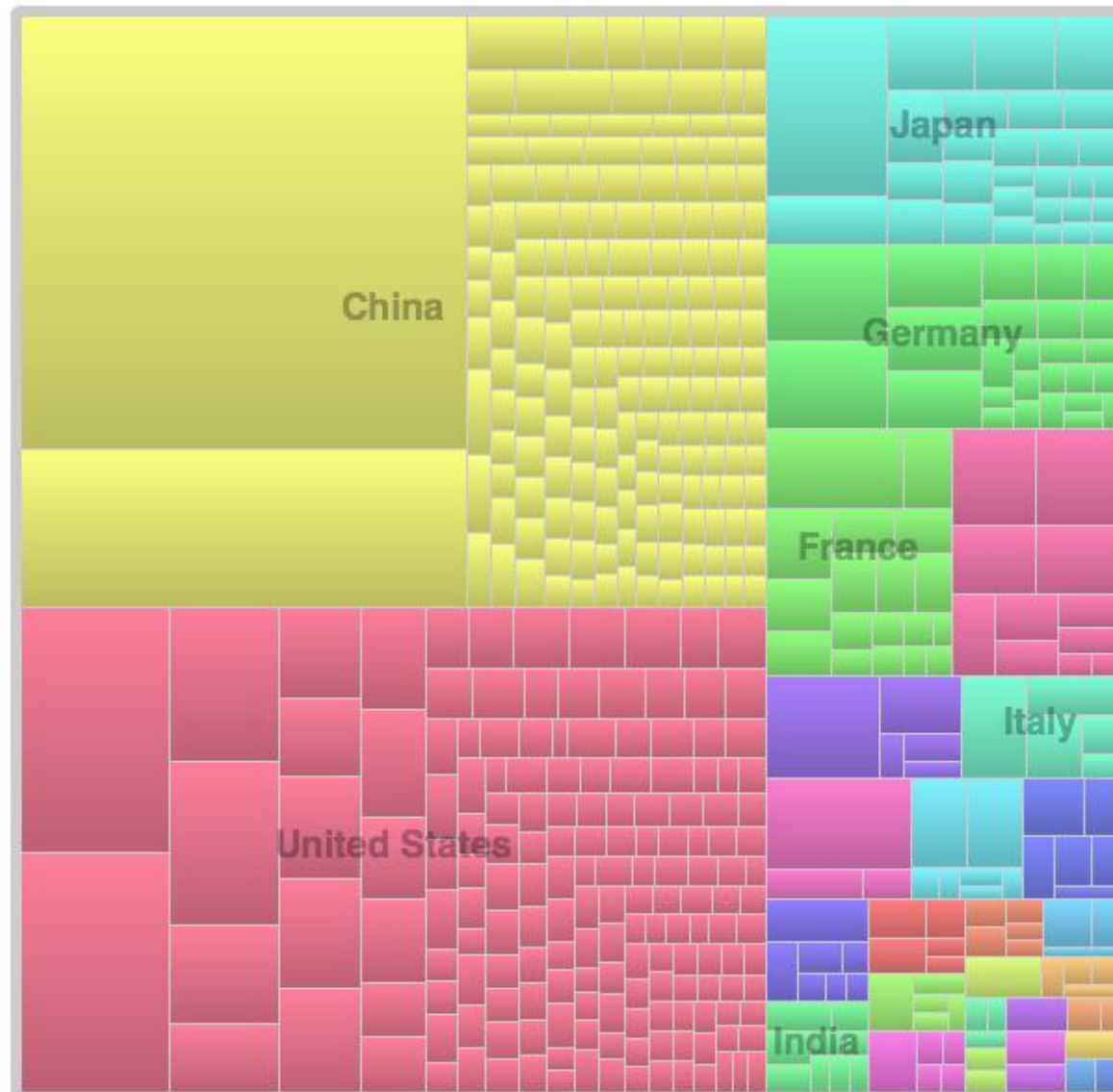
COUNTRY	NUMBER OF SUPERCOMPUTERS	
China	**167**	
United States	165	
Japan	29	
Germany	26	
France	18	
Britain	12	
India	9	
Russia	7	
South Korea	7	
Poland	6	
other	54	

China has 1/3 of the systems, while the number of systems in the US has fallen to the lowest point since the TOP500 list was created.

COUNTRIES
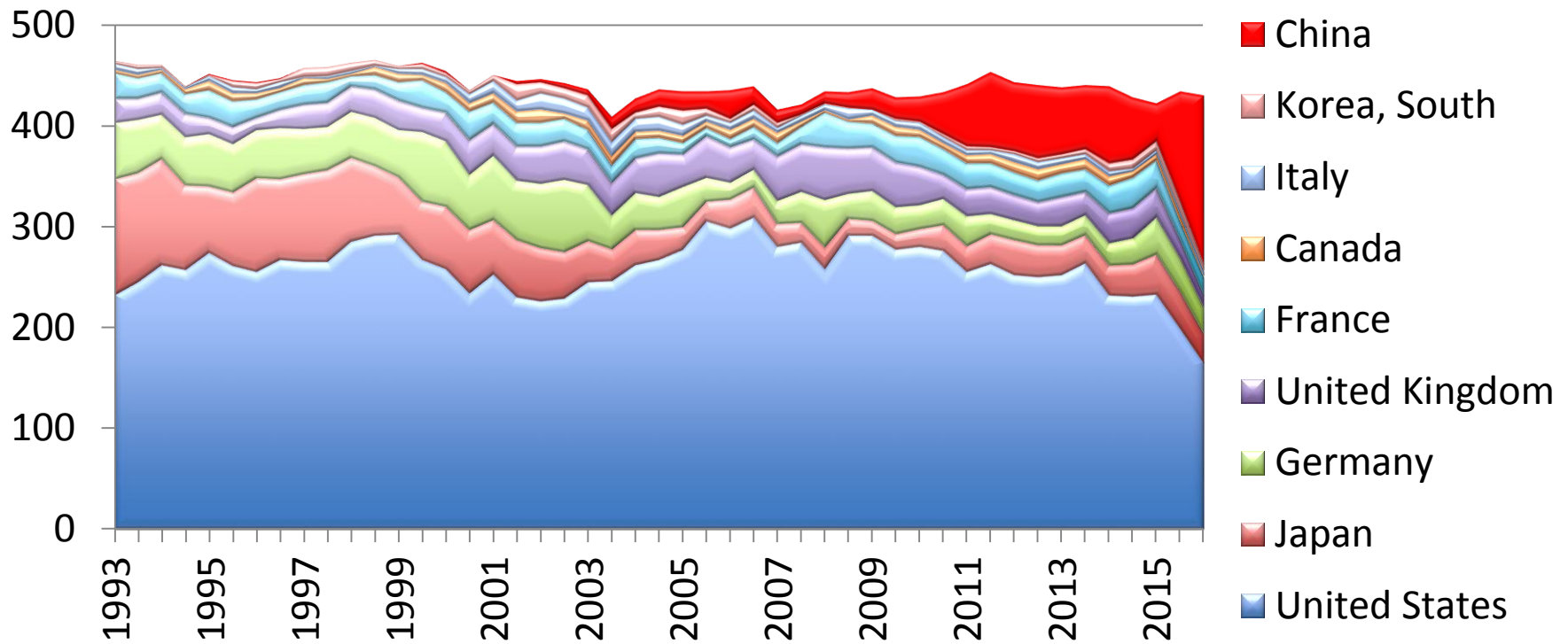
Performance of Countries
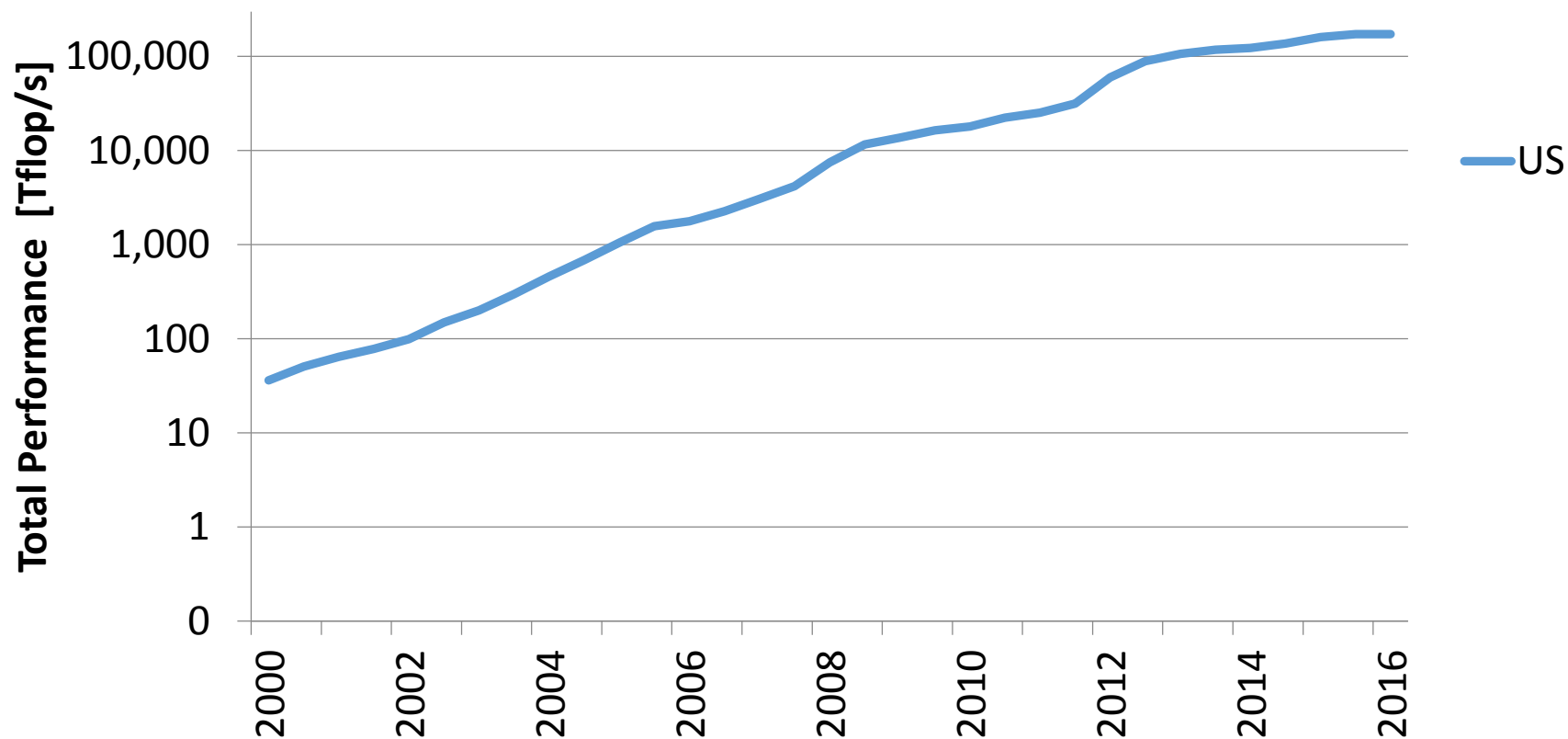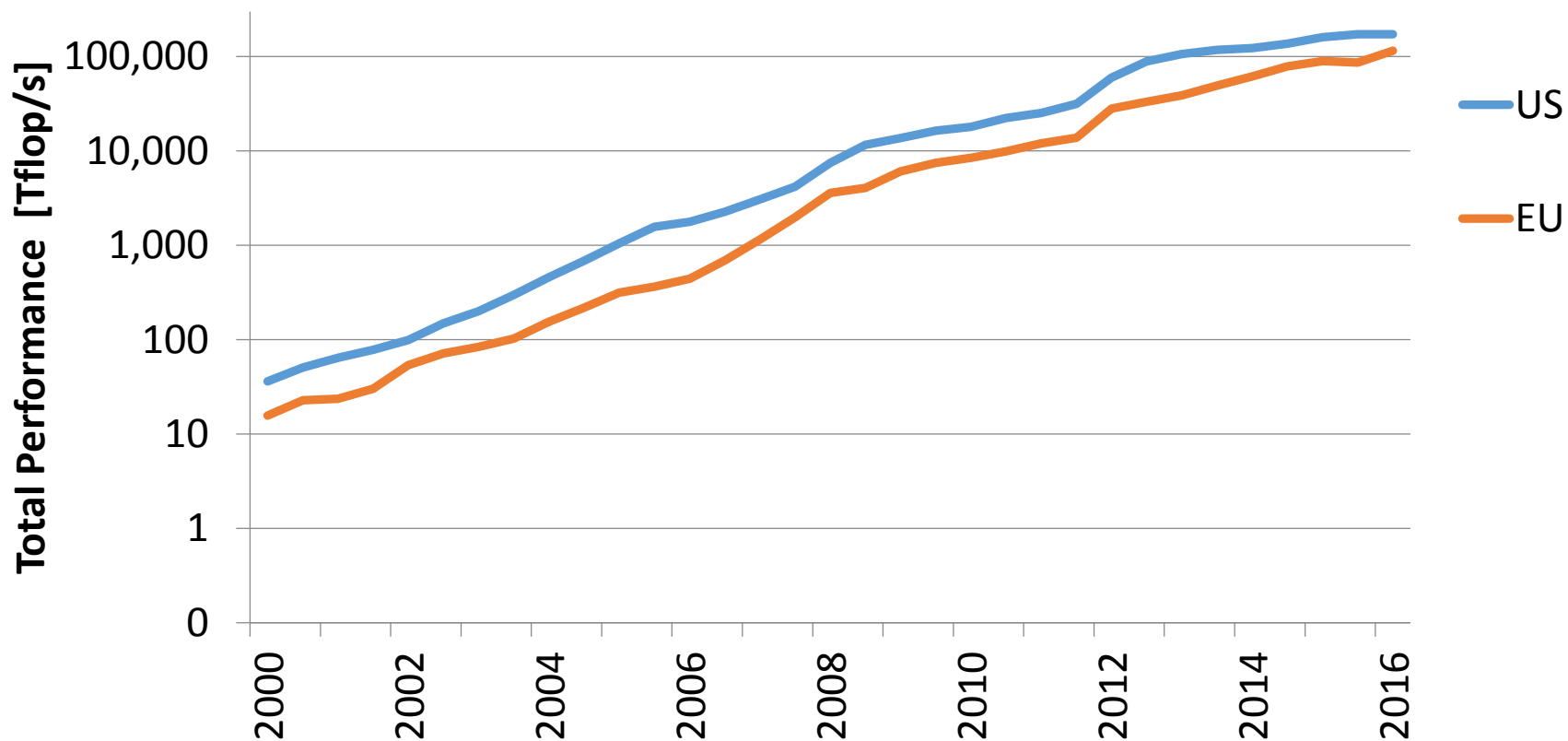
Performance of Countries

Performance of Countries

Performance of Countries

Total Performance [Tflop/s] vs Year chart showing US, EU, Japan, and China from 2000 to 2016.

The average age of systems:
US grew from 1.25 to 2.25
Europe to 2.75
Japan always had 2-3 years old systems, now close to 3 years
China is at 1.6 years and has youngest population by far

Vendors / System Share

IBM/Lenovo, 8, 2%
NUDT, 4, 1%
Others, 41, 8%
Dell, 9, 2%
Fujitsu, 10, 2%
Inspur, 20, 4%
Bull, 22, 4%
SGI, 26, 5%
IBM, 38, 8%
Sugon, 51, 10%
HP, 127, 25%
Lenovo, 84, 17%
Cray Inc., 60, 12%

- HP
- Lenovo
- Cray Inc.
- Sugon
- IBM
- SGI
- Bull
- Inspur
- Fujitsu
- Dell
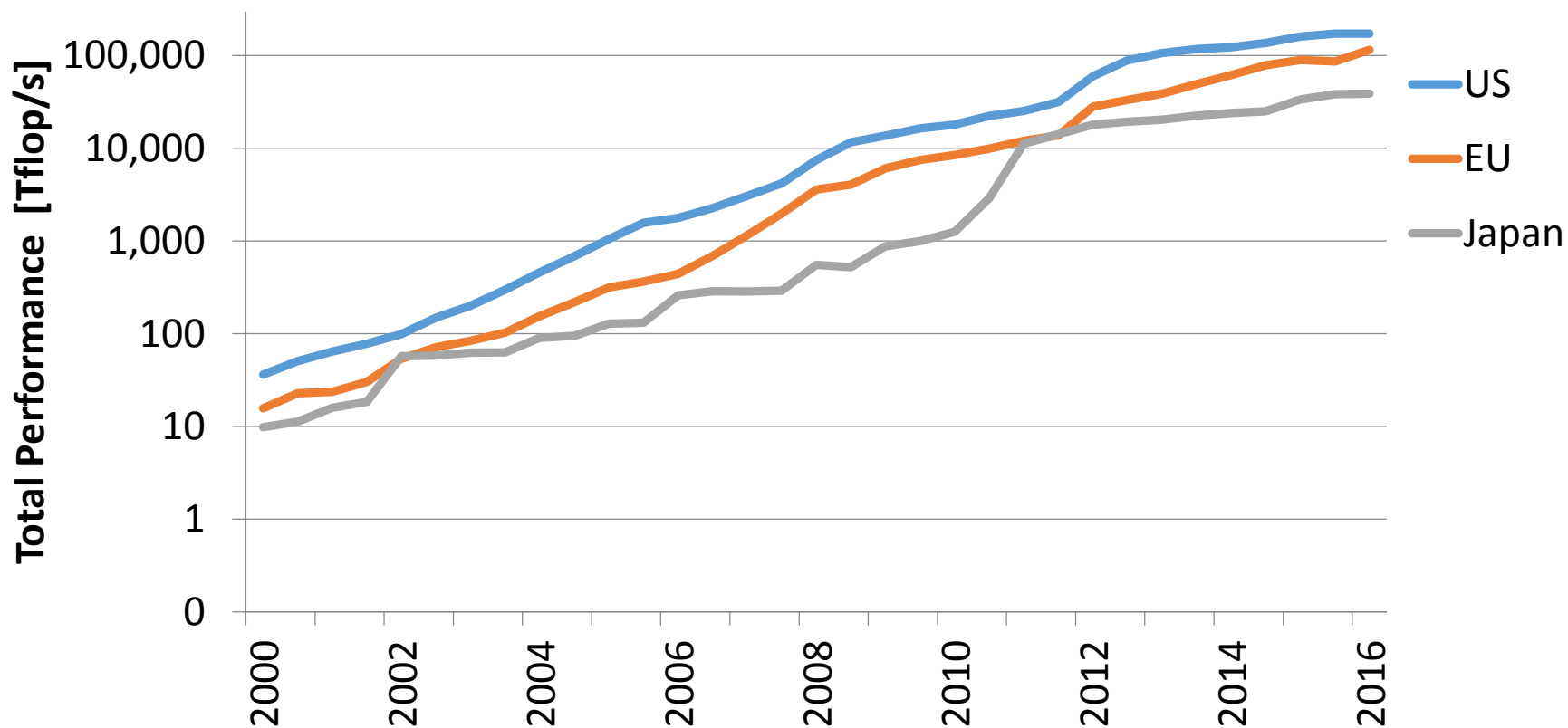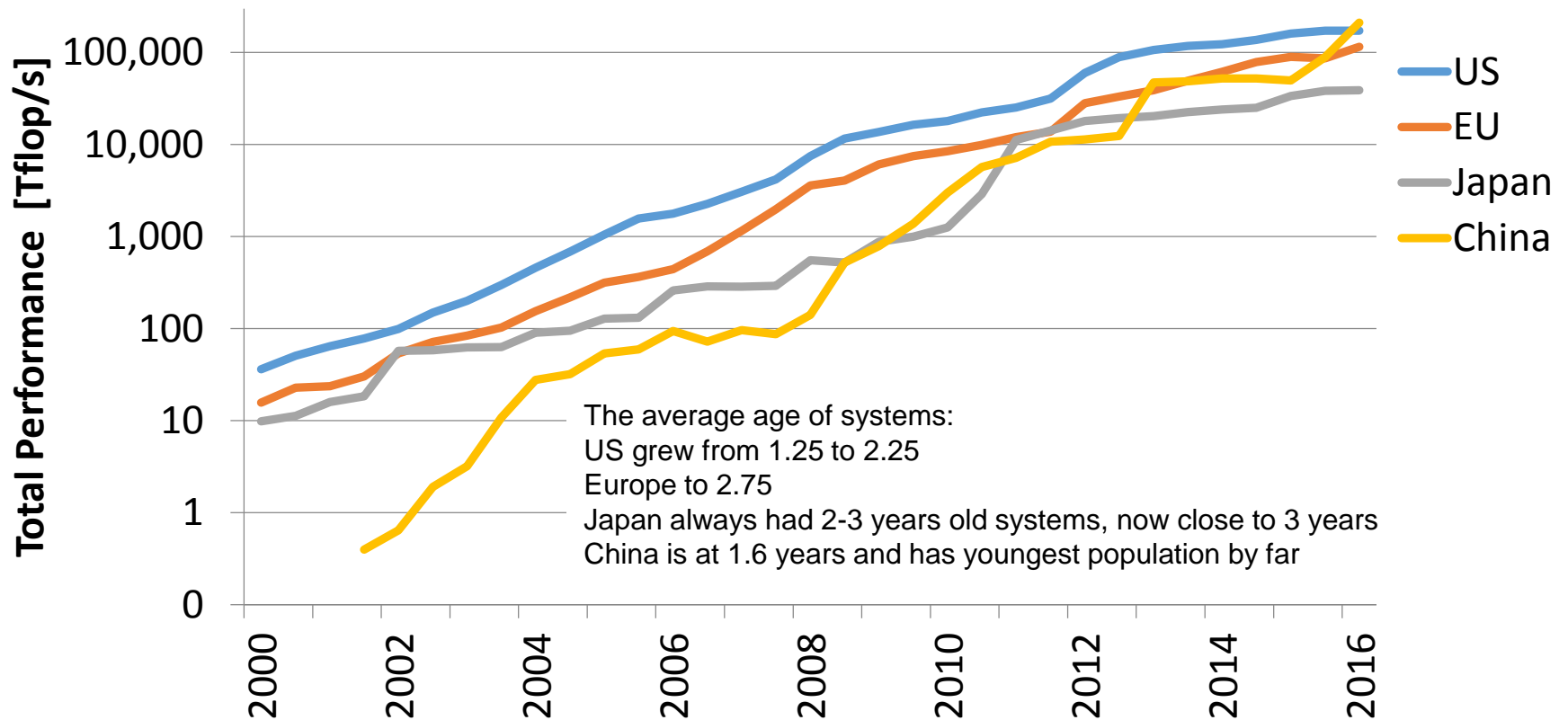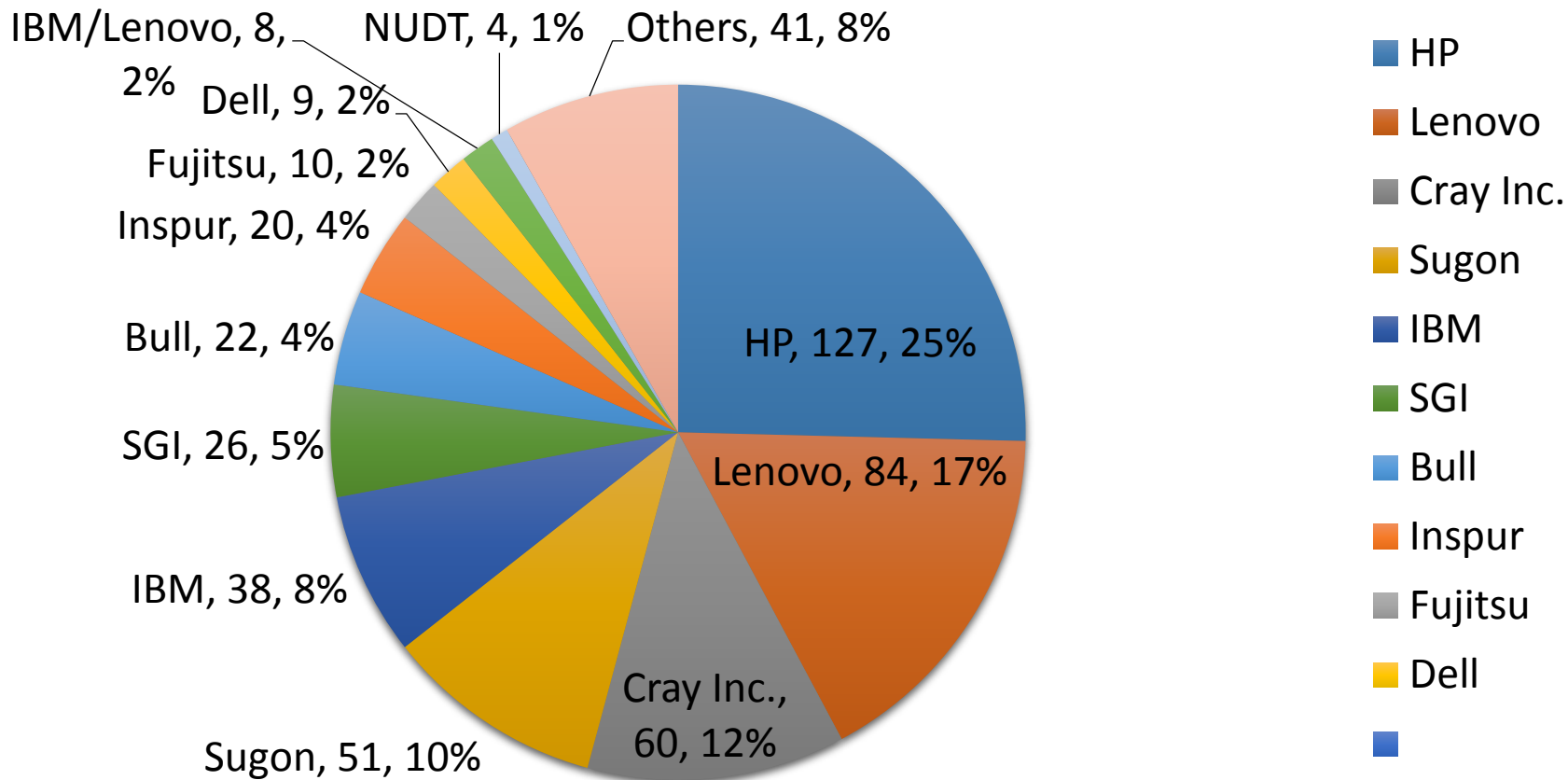
Vendor, # of systems, % of Top500

Vendors / Performance Share

Others, 27, 5% HP, 58, 10%
Lenovo, 44, 8%
NRCPC, 94, 16%
NUDT, 39, 7%
IBM/Lenovo, 10, 2%
Cray Inc., 113, 20%
Dell, 12, 2%
Fujitsu, 24, 4%
Inspur, 9, 1%
Bull, 22, 4%
SGI, 32, 6% IBM, 61, 11%
Sugon, 23, 4%

Legend:
- HP
- Lenovo
- Cray Inc.
- Sugon
- IBM
- SGI
- Bull
- Inspur
- Fujitsu
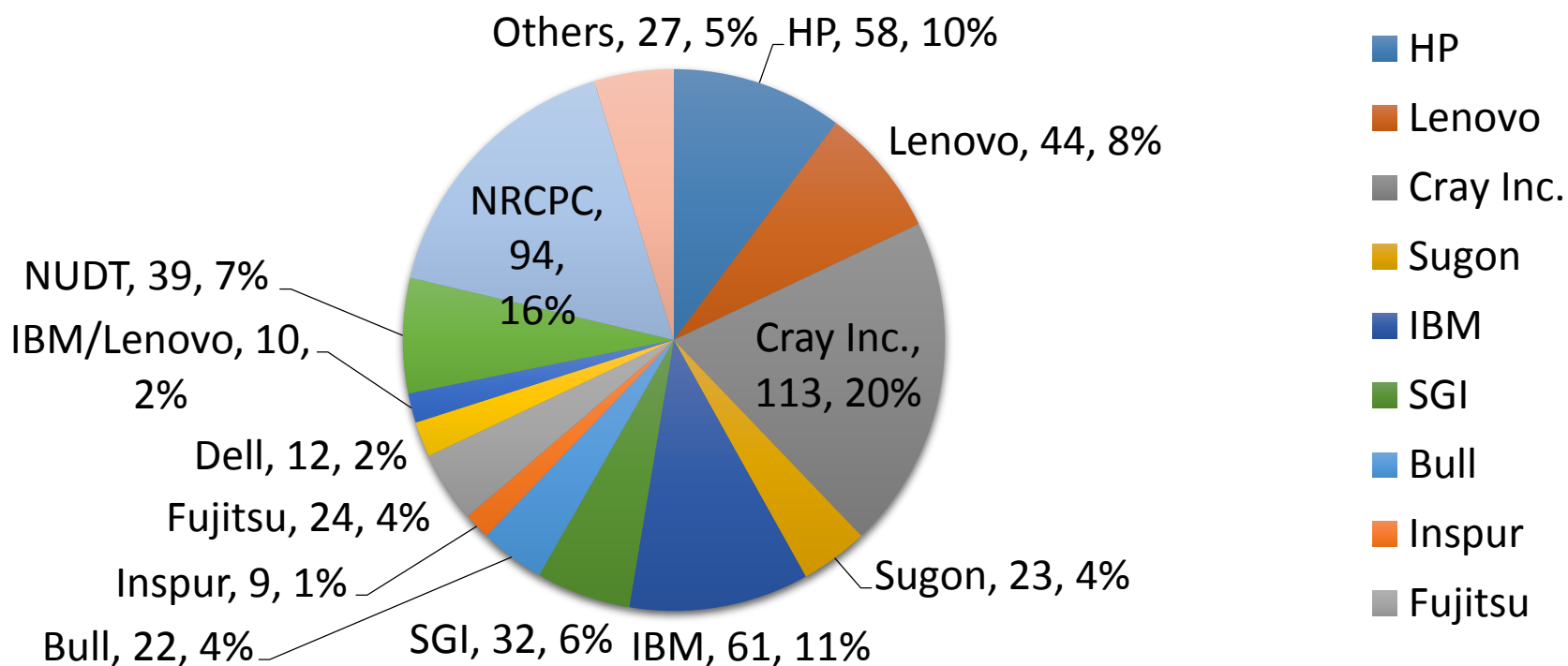
Vendor, Sum of Pflop/s, % of Top500

Sunway TaihuLight, was developed at the National Research Center of Parallel Computer Engineering and Technology (NRCPC)

2013 CHINA TOP 100
(major US presence)

Dawning name change to Sugon

On-the-ground in Asia
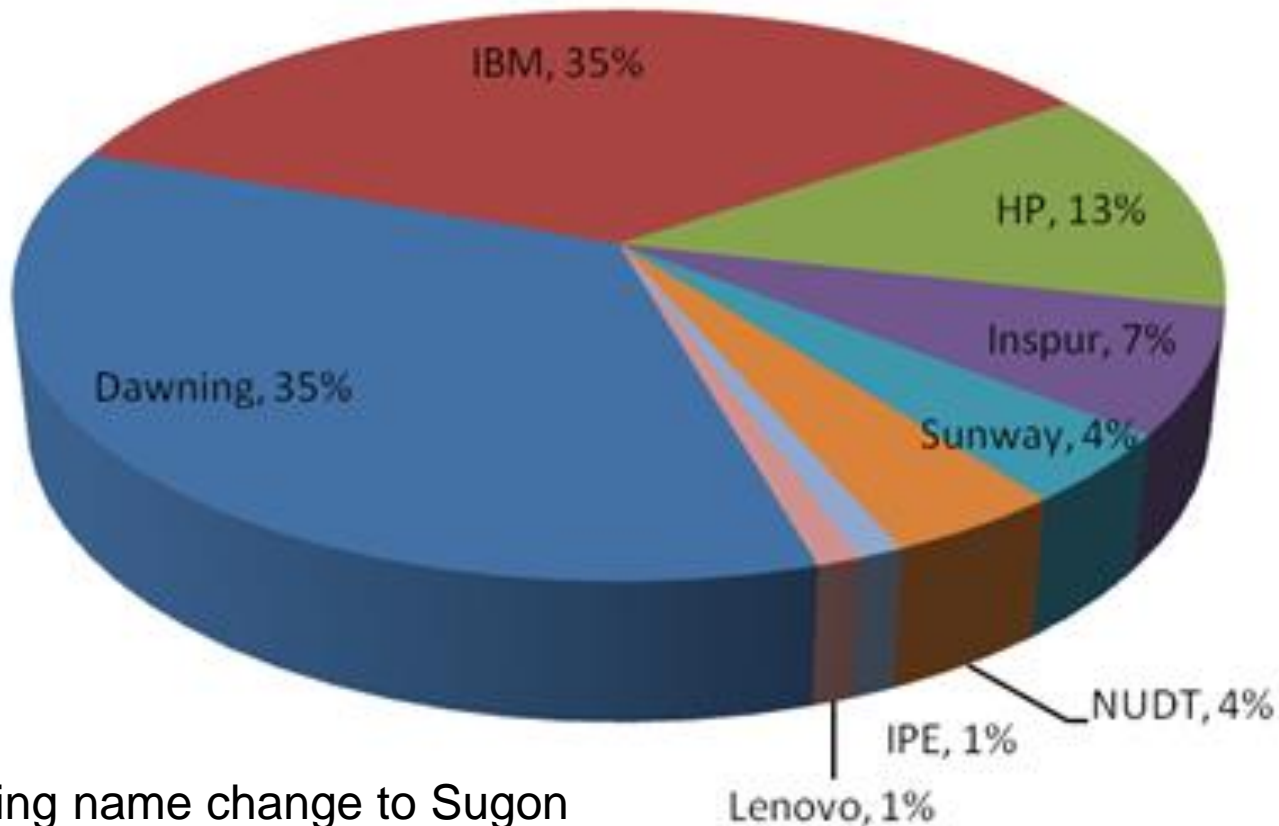
2014 2015 Chinese Top100
(US is nearly eliminated)

2014 Market Share of Top100 HPC Systems

- NRCPCET 1%
- IPE CAS 1%
- Huawei 1%
- IBM 3%
- NUDT 4%
- HP 5%
- Dawning 32%
- Inspur 21%
- Lenovo 32%
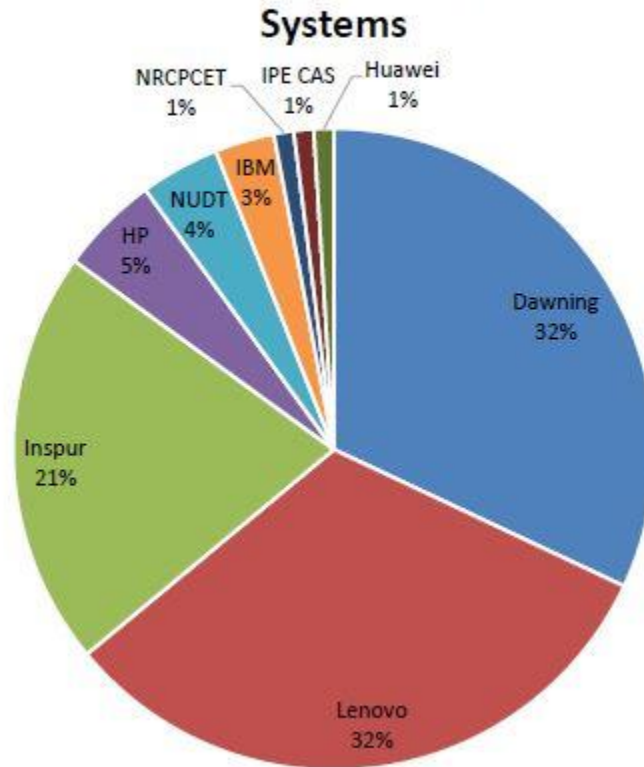
Foreign/Domestic Divide by Year

Foreign — Domestic
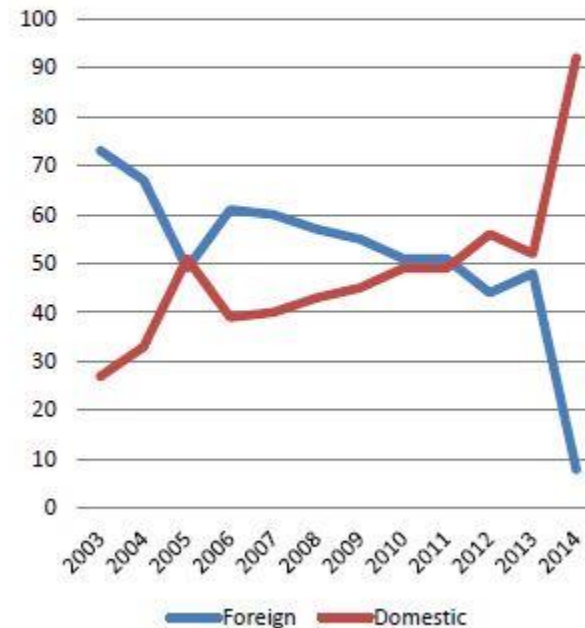
➔**2015: Sugon (Dawning) 34; Lenovo 34; Inspur 23; Total 91% (IBM 2; HP 1)**

atip
asian **technology**
information program

On-the-ground in Asia

System with HPL > 1 Pflop/s
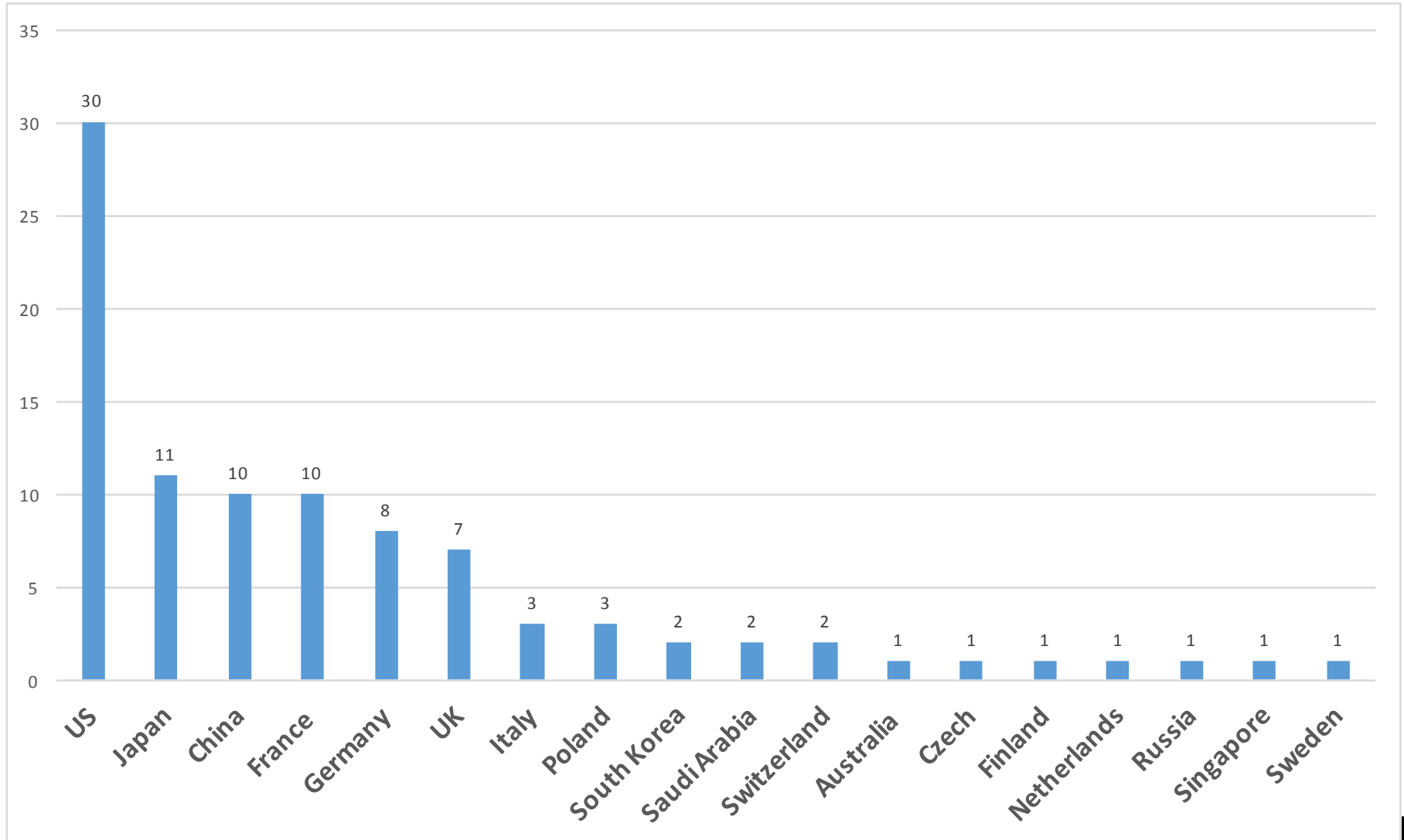
Recent Developments

- US DOE planning to deploy O(100) Pflop/s systems for 2017-2018 - $525M hardware
- Oak Ridge Lab and Lawrence Livermore Lab to receive IBM, Nvidia, Mellanox based systems
- Argonne Lab to receive Intel and Cray based system
  - After this the next round of systems are an Exaflop
- US Dept of Commerce is groups from receiving Int
  - cle th the
  - ngzh it f Ti l
  - njin,
  - National SC Center Changsh

Since the Dept of Commerce Action …

- **Expanded focus on Chinese made HW and SW**
  - **"Anything but from the US"**
- **Three separate developments in HPC**
  - **Jiangnan**
    - **ShenWei O(100) Pflops all Chinese, June 2016**
  - **NUDT**
    - **Tianhe-2A O(100) Pflops will be Chinese ARM + accelerator, 2017**
  - **Sugon - CAS ICT**
    - **AMD? new processors**
- **In the latest "5 Year Plan"**
  - **Govt push to build out a domestic HPC ecosystem.**
  - **Exascale system, will not use any US chips**

9/20/2016

Rapid development of supercomputers

➤ Top 1 machines build by China.

Top 1 of 2010.11
Tianhe-1A
- Peak: 4.7 PF
- HPL: 2.5 PF
- Cores: 186 K
- Arch: CPU-GPU

➤ Main trend observed

- An abrupt performance jump at each announcement
- A variety of heterogeneous archs: CPU-GPU, CPU-MIC, and beyond ...
- Number of cores grows from O(100K) to O(10M) in only 6 years

9/20/2016

Rapid development of supercomputers

➤ Top 1 machines build by China.

Top 1 of 2010.11
Tianhe-1A
- Peak: 4.7 PF
- HPL: 2.5 PF
- Cores: 186 K
- Arch: CPU-GPU

Top 1 of 2013.06
Tianhe-2
- Peak: 55 PF
- HPL: 33.8 PF
- Cores: 3.12 M
- Arch: CPU-MIC

➤ Main trend observed

- An abrupt performance jump at each announcement
- A variety of heterogeneous archs: CPU-GPU, CPU-MIC, and beyond …
- Number of cores grows from O(100K) to O(10M) in only 6 years

9/20/2016

Rapid development of supercomputers

➢ Top 1 machines build by China.

Top 1 of 2010.11
Tianhe-1A
- Peak: 4.7 PF
- HPL: 2.5 PF
- Cores: 186 K
- Arch: CPU-GPU

Top 1 of 2013.06
Tianhe-2
- Peak: 55 PF
- HPL: 33.8 PF
- Cores: 3.12 M
- Arch: CPU-MIC

Top 1 of 2016.06
Sunway TL
- Peak: 125 PF
- HPL: 93 PF
- Cores: 10.5M
- Arch: SW CPU

➢ Main trend observed

- An abrupt performance jump at each announcement
- A variety of heterogeneous archs: CPU-GPU, CPU-MIC, and beyond ...
- Number of cores grows from O(100K) to O(10M) in only 6 years

9/20/2016

The Sunway Series Supercomputers

Sunway BlueLight
#120 on Top500

- JinanNational Supercomputing Center
- 1 PFlops
- multi-core processor (16-core)
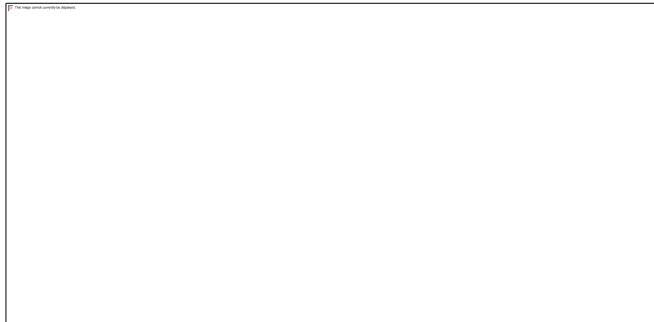
Sunway TaihuLight
#1 on Top500

- National Supercomputing Center in Wuxi
- 125 PFlops
- many-core processor (260-core)

Sunway TaihuLight

- System vendor: National Research Center Of Parallel Computer Engineering & Technology （NRCPC）

- CPU vendor: Shanghai High Performance IC Design Center

- Facility host: National Supercomputing Center in Wuxi, a joint team by Tsinghua University, City of Wuxi, and Jiangsu Province

SW26010 Procesor

- China's first homegrown many-core processor
  - Vendor: Shanghai High Performance IC Design Center
  - supported by National Science and Technology Major Project (NMP): Core Electronic Devices, High-end Generic Chips, and Basic Software
  - 28 nm technology
  - 260 Cores
  - 3 Tflop/s peak

SW26010: General Architecture

- 1 node

- 260 cores per processor

- 4 Core Groups (CGs), each of wh

  - 1 Management Processing Elemer

  - 64 (8x8) Computing Processing El

RoadRunner Node

Roadrunner Node

Memory

Cell

Dual-Core Microprocessor   Single Core

Parallel Computing Network (Infiniband DDR)

Main memory				Main memory				Main memory				Main memory		
MC	Master core	MPE		MC	Master core	MPE		MC	Master core	MPE		MC	Master core	MPE	

Slave cores

64k

64KB
cache/core

Group

Slave cores

64k

64KB
cache/core

Group

Slave cores

64k

64KB
cache/core

Group

Slave cores

64k

64KB
cache/core

Group

Network on Chip (NoC)

SI

SW26010: MPE and CPE

- Management Processing Element (MPE)
  - 64-bit RISC core
  - support both user and system modes
  - 256-bit vector instructions
  - 32 KB L1 instruction cache, and 32 KB L1 data cache
  - 256 KB L2 cache
- Computing Processing Element (CPE)
  - 8x8 CPE mesh
  - 64-bit RISC core
  - support only user mode
  - 256-bit vector instructions
  - 16 KB L1 instruction cache, and 64 KB Scratch Pad Memory (total for the 64 CPE)

FEATURE	INTEL® XEON PHI™ COPROCESSOR 7120P	Intel® Xeon Phi™ Processor (codename Knights Landing)	Sunway TaihuLight Node
Processor Cores	Up to 61 enhanced P54C Cores	Up to 72 enhanced Silvermont cores	260 cores / node
Key Core Features	In order, 4 threads / core, 2 wide	Out of order, 4 threads / core, 2 wide	1 thread / core
High Speed Memory	Up to 16 32-bit channels GDDR5 @ up to 5.5GT/s	Eight 128-bit channels MCDRAM @ 7.2 GT/s	Up to 4 128-bit channels
Off Package Memory	None	6 channels DDR4 2400MHz	4*128 channels DDR3 at 2133 MHz
Memory Bandwidth	Up to 181 GB/s STREAM Triad (GDDR5)	~ 490 GB/s STREAM Triad (to MCDRAM) + ~ 90GB/s STREAM Triad (to DDR4)	136 GB/s *128-bit DDR3-2133*
Memory Capacity	Up to 16 GB on-package GDDR5	16 GB on package memory (MCDRAM) + Up to 384 GB off package DDR4	32 GB off package DDR3
Peak FLOPS	SP: 2.416 TFLOPs; DP: 1.208 TFLOPs	Up to SP 6.912 TFs (at 1.5GHz TDP freq) Up to DP 3.456 TFs (at 1.5GHz TDP freq)	DP: 3.06 Tflop/s SP: Same as DP
FLOPS/Byte (from memory)	1.208 Tflop/s / 181 GB/s = 6.67 Flops/Byte	3.456 TFLOP/s at 490 GB/s = 7.05 Flops/Byte	3.06 Tflop/s / 136.51 GB/s = 22.4 Flops/Byte

Sunway TaihuLight http://bit.ly/sunway-2016

- SW26010 processor
- Chinese design, fab, and ISA
- 1.45 GHz
- Node = 260 Cores (1 socket)
  - 4 – core groups
    - 64 CPE, No cache, 64 KB scratchpad/CG
    - 1 MPE w/32 KB L1 dcache & 256KB L2 cache
  - 32 GB memory total, 136.5 GB/s
  - ~3 Tflop/s, (22 flops/byte)
- Cabinet = 1024 nodes
  - 4 supernodes=32 boards(4 cards/b(2 no
  - ~3.14 Pflop/s
- 40 Cabinets in system
  - 40,960 nodes total
  - 125 Pflop/s total peak
- 10,649,600 cores total
- 1.31 PB of primary memory (DDR3)
- 93 Pflop/s for HPL, 74% peak
- 0.32 Pflop/s for HPCG, 0.3% peak
- 15.3 MW, water cooled
  - 6.07 Gflop/s per Watt
- 3 of the 6 finalists Gordon Bell Award@SC16
- 1.8B RMBs ~ $280M, (building, hw, apps, sw, …)

CPE cluster

Main Memory   Main Memory
MC   CPE Cluster   MC   CPE Cluster
MPE   CG   MPE   CG

SI   Network on Chip (NoC)

L1   MPE   CPE Cluster   MPE   CPE Cluster
L2   MC   CG   MC   CG

Main Memory   Main Memory

SPM

8 × 8

Cabinet (4 Supernodes)   Cabinet (4 Supernodes)   Cabinet (4 Supernodes)

40 Cabinets
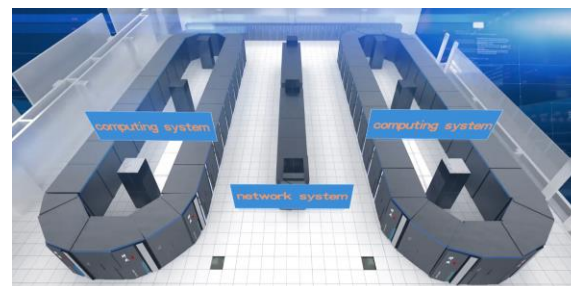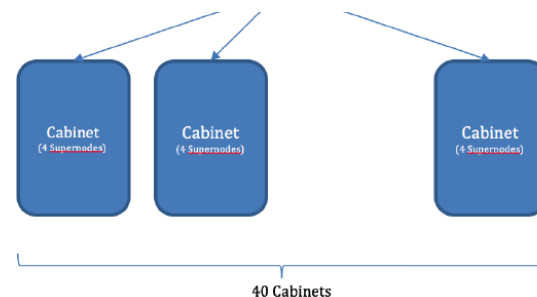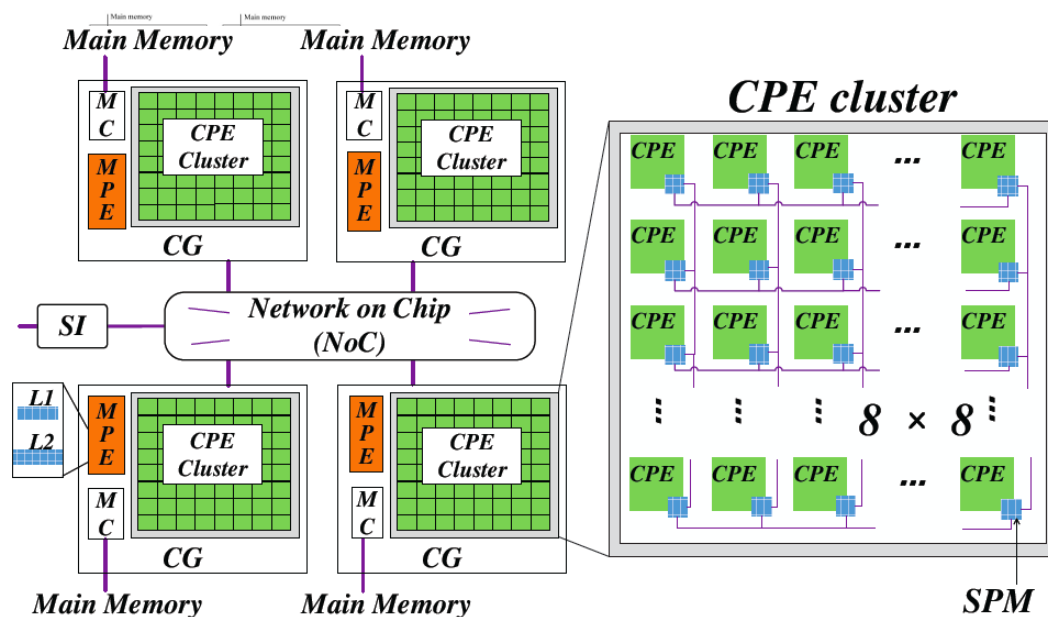
Socket = 1 Node = 260 Cores, 3.06 Tflop/s
Card = 2 Nodes = 6.12 Tflop/s
Board = 4 Cards = 24.48 Tflop/s
Board is water cooled
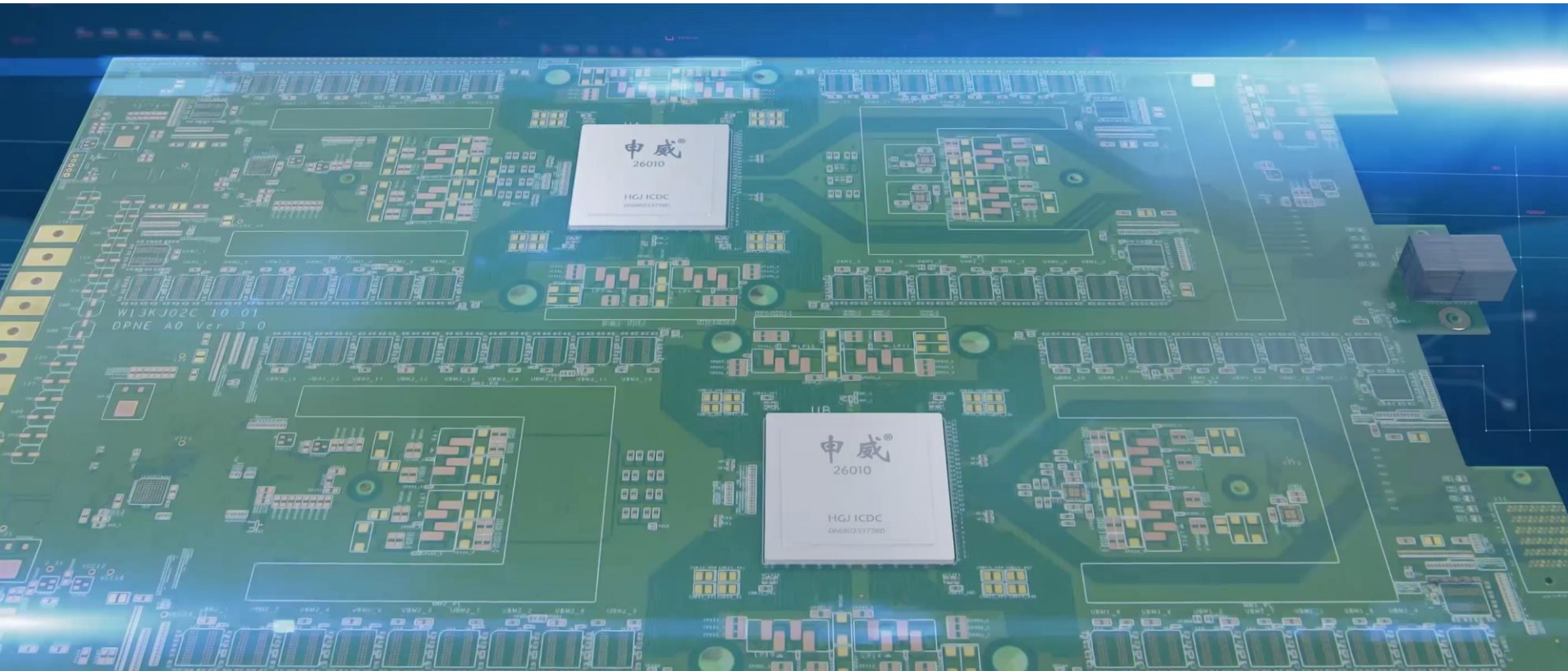
Supernode = 32 Boards
= 256 Nodes
= 783.97 Tflop/s
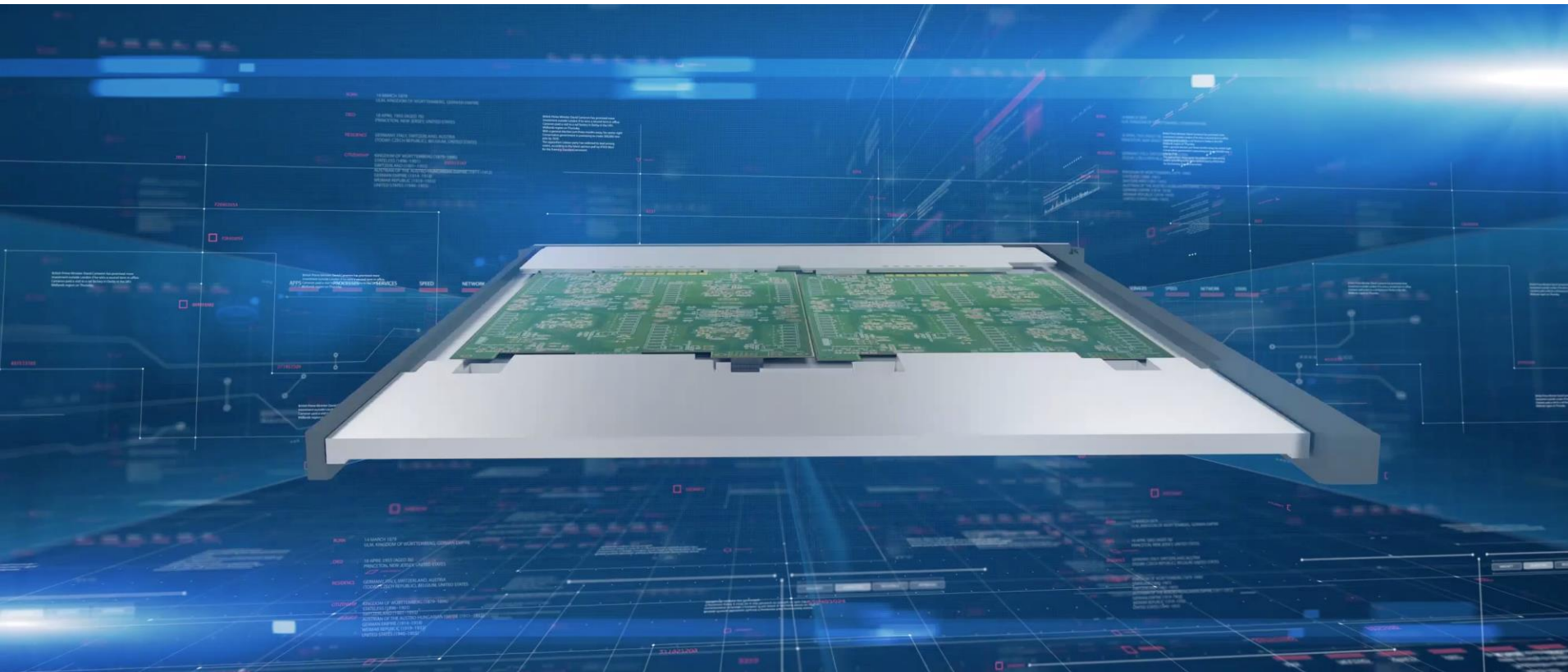
Cabinet = 4 Supernodes = 3.1359 Pflop/s

High density integration of the reconfigurable super node architecture

System = 40 Cabinets = 160 Supernodes
= 40,960 Nodes = 10,649,600 cores
= 125.4359 Pflop/s

Software System

Parallel Application

Parallel Program Development Environment
- IDE
- Parallel Debug
- Performance Monitor

Parallel Compiling Environment
- ●OpenACC
- ●OpenMP
- ●MPI

Many-Core Basic Software

Compiling System
- ● C/C++、Fortran
- ● SIMD

Basic Libs
- ●C Lib
- ●ACC Thread Lib
- ●Math Lib

Auto-vectorization
- ●C、C++、Fortran
- ●Loop Vectorization
- ●Code Optimization

Parallel OS Environment
- ●Job
- ●Resource
- ●Power
- ●Network
- ●Fault Tolerant
- ●System
- ●Security

HPC Storage Management
- ●SWGFS
- ●LWFS
- ●Storage Management Platform

Sunway TaihuLight System

➢ **xMath: a high-performance extended math library**

- Compatible with commercial libs
  - MKL, ACML, ESSL, etc
- Four major modules
  - BLAS, LAPACK, FFT, Sparse
- Tailored for the Sunway CPU
  - Assemble-level optimization
  - Instruction-level optimization
  - Data-movement reduction by LDM/DMA
  - Pthreads parallelism for MPEs
  - Athread parallelism for CPEs
- Version history
  - V0.9a released on 01/01/2016
  - V1.0b released on 06/20/2016
- We use xMath for optimized BLAS1 ops.
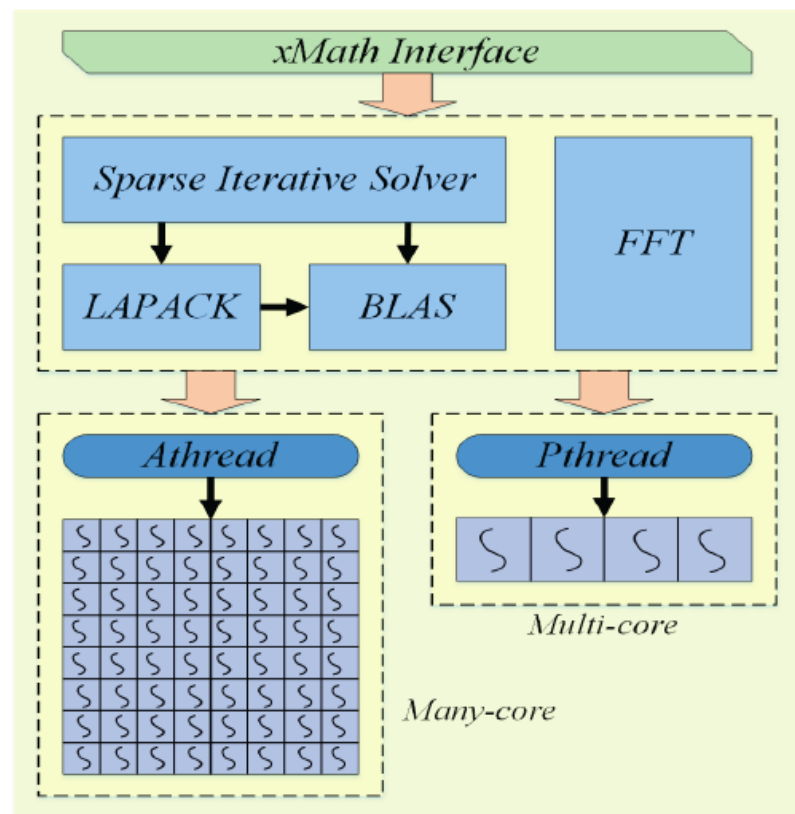
Sunway TaihuLight Storage System

Total Capacity	Over 10 PB
Peak I/O Throughput	Over 288 GB / s
# of Storage Nodes	288
# of Disk Arrays	144
# of SSD	288
Capacity per SSD	800 GB

Business Software
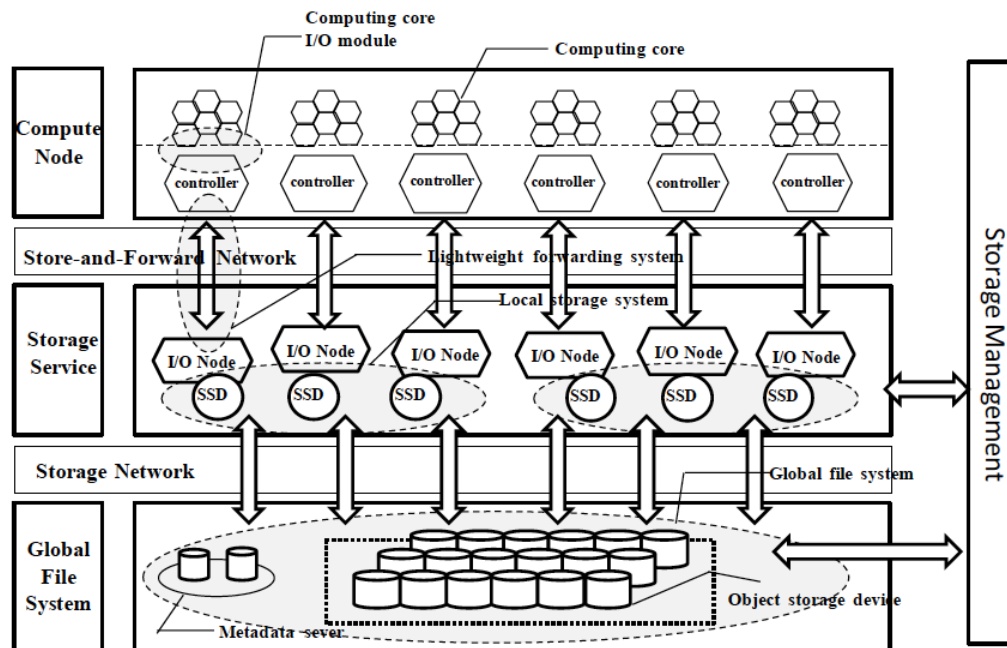
Software Type	Software Title	Software Description
Solver	ANSYS.Fluent	Computational Fluid Dynamics Analysis Software
	ANSYS.CFX	Computational Fluid Dynamics Analysis Software
	STAR-CCM+	Computational Fluid Dynamics Analysis Software
	ABAQUS	Computational Structure Dynamics Analysis Software
	LS-DYNA	Computational Structure Dynamics Analysis Software
	Radioss	Computational Structure Dynamics Analysis Software
	ANSYS.Mechanical	Computational Structure Dynamics Analysis Software
	MSC.NASTRAN	Computational Structure Dynamics Analysis Software
	FEKO	High Electromagnetic Field Analysis Software
	OptiStruct	Structural Mechanics Analysis and Optimization Software
Pre/Post processing	MSC.PATRAN	FEM Pre/Post Processing Software
	HyperMesh	FEM Pre/Post Processing Software
	ICEM-CFD	CFD Pre/Post Processing Software
	GAMBIT	Computational Fluid Pre-Processing Software

Scientific Computing Software

Software Title	Software Description
WRF	Mesoscale Weather Forecast Model
OpenFOAM	Fluid Dynamics Software
Gaussian	Hyperchem
ABINIT	Computational Chemistry Software
NWChem	Computational Chemistry Software
Q-Chem	Computational Chemistry Software
CP2K	Material Science Software
CPMD	Material Science Software
VASP	Material Science Software
AMBER	Computational Biology Software
GROMACS	Computational Biology Software
LAMMPS	Computational Biology Software
NAMD	Computational Biology Software
DOCK	Molecular Docking Program
BLAST	Column Similarity Search Program

Sunway TaihuLight: Applications

Key application domains:

Earth system modeling / weather forecasting

Advanced manufacturing (CFD/CAE)

Life science

Big data analytics

国家超级计算无锡中心
National Supercomputer Center in Wuxi

Applications on the TaihuLight

- 35 apps running on the system
  - 6 of them are running at full scale
  - 18 of them are running on half the machine
  - 20 applications on million cores
- Apps will typically run "out of the box"
  - No use of CPEs, just on MPE, with poor performance
  - Codes needs to be refactored to use CPE
- The Center has 20 people to help with optimizing apps to run on the system.
- CAM code 20K lines of code to start, ended with 100K lines, 10 people.
- Phase field 12K lines of code to start, ended with 20K, 3 people + help

Our mini co-design: a tale of 4 geeks

Application — Earth System Modeling

Algorithm — Dynamic Core Solver

Architecture — Implementation & Optimization
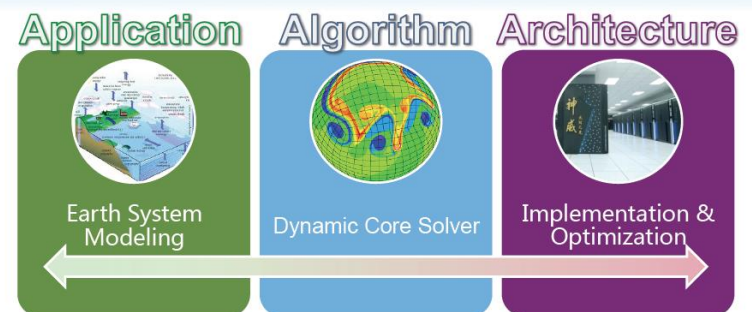
Prof. Lanning Wang (BNU)

Prof. Chao Yang (ISCAS)

Prof. Wei Xue (Tsinghua U)
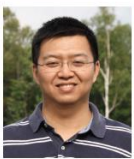
Prof. Haohuan Fu (Tsinghua U) (NSCC-Wuxi)
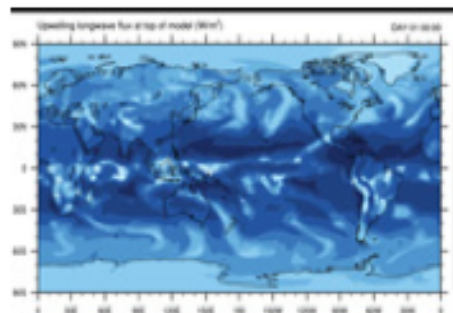
CAM : Community Atmosphere Model

Refactoring and Optimizing the CAM on the New Sunway Many-core Supercomputer ...
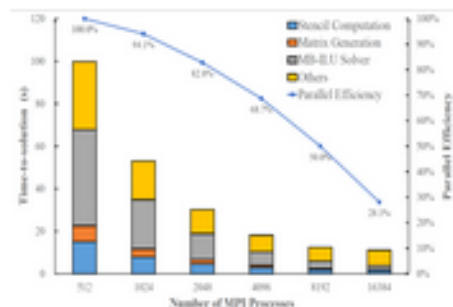
Nonhydrostatic Atmospheric Dynamics

Peta-Scale Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics with 8.5M Cores ...
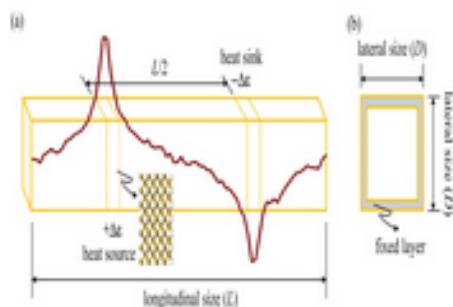
Silicon Nanowires

Large-scale Molecular Dynamics Simulation of Thermal Conductivity of Silicon Nanowires. ...

3D Parallel Numerical Simulation

3D Parallel Numerical Simulation for Seismic Forward Modeling

...
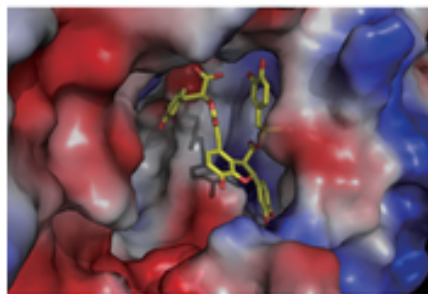
Drugs

Binding Affinity Prediction of Drugs and Molecular Dynamic Simulations of Biomacromolecules ...
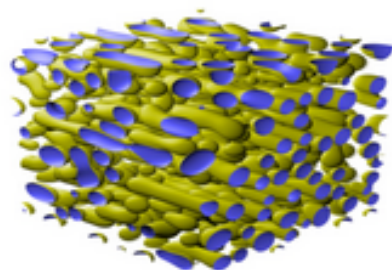
Coarsening Dynamics

Large Scale Phase Field Simulation for Coarsening Dynamics Based on Cahn-Hilliard Equation with Degenerated Mobility ...
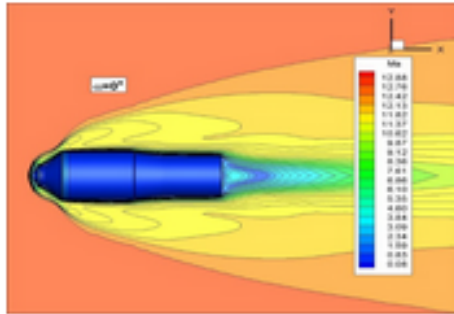
The Aerospace-craft Unification Algorithm

Numerical Simulation of the Aerospace-craft Unification Algorithm ...
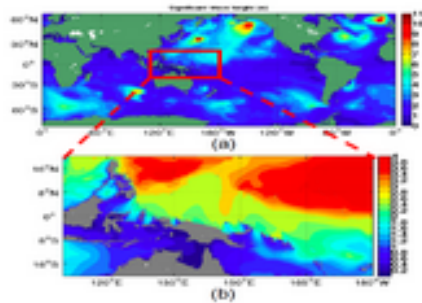
Global Surface Wave Numerical Simulation

A Highly Effective Global Surface Wave Numercial Simulation with Ultra-high Resolution ...

For SC16
Five Gordon Bell Submissions From China running on the TaihuLight System

A Fully-Implicit Nonhydrostatic Dynamic Solver for Cloud-Resolving Atmospheric Simulation

A highly effective global surface wave numerical simulation with ultra-high resolution

Large Scale Phase-Field Simulation for Coarsening Dynamics

Peta-scale Atomistic Simulation of Silicon Nanowires

Time-parallel molecular dynamics simulations

Gordon Bell Finalists at SC16

- "**Modeling Dilute Solutions Using First-Principles Molecular Dynamics: Computing More than a Million Atoms with Over a Million Cores**,"
  - Lawrence-Livermore National Laboratory (Calif.)
- "**Towards Green Aviation with Python at Petascale**,"
  - Imperial College London (England)
- "**Simulations of Below-Ground Dynamics of Fungi: 1.184 Pflops Attained by Automated Generation and Autotuning of Temporal Blocking Codes**,"
  - RIKEN (Japan), Chiba University (Japan), Kobe University (Japan) and Fujitsu Ltd. (Japan)
- "**Extreme-Scale Phase Field Simulations of Coarsening Dynamics on the Sunway Taihulight Supercomputer**,"
  - Chinese Academy of Sciences, the University of South Carolina, Columbia University (New York), the National Research Center of Parallel Computer Engineering and Technology (China) and the National Supercomputing Center in Wuxi (China)
- "**A Highly Effective Global Surface Wave Numerical Simulation with Ultra-High Resolution**,"
  - First Institute of Oceanography (China), National Research Center of Parallel Computer Engineering and Technology (China) and Tsinghua University (China)
- "**10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics**,"
  - Chinese Academy of Sciences, Tsinghua University (China), the National Research Center of Parallel Computer Engineering and Technology (China) and Beijing Normal University (China)

Sunway TaihuLight is Available …

- The TaihuLight was put on the internet last week.
- If you would like to use the TaihuLight, go to…
  - http://www.nsccwx.cn/wxcyw/process.php?word=process&i=54

Key Points Re Chinese HPC

Expanded focus on Chinese made HW and SW
    Anything but from the US

Three separate threads developing HPC
    Jiangnan (Shenwei), NUDT (Tianhe), CAS-ICT (Godson/Sugon)

Govt push to build out a domestic HPC ecosystem

Coupling this with indigenous semiconductor development and
    acquisition

a t i p
asian **technology**
information program

Some Plans

- Planning a smaller version of the system
  - One rack maybe by the end of the year
  - Air cooled version
- Sunway is interested in marketing the system

Worldwide

- Quickly look at
  - Japan
  - Europe

An Overview of Flagship 2020 project

- **Developing the next Japanese flagship computer, so-called "post K"**

Vendor partner

- **Developing a wide range of application codes, to run on the "post K", to solve major social and science issues**

The Japanese government selected 9 social & scientific priority issues and their R&D organizations.

FUJITSU

Society with health and longevity	Disaster prevention and global climate	Energy issues	Industrial competitiveness

①Innovative Drug Discovery
RIKEN Quant. Biology Center

③Hazard and Disaster induced by Earthquake and Tsunami
Earthquake Res. Inst., U. Tokyo

⑥Innovative Clean Energy Systems
Grad. Sch. Engineering, U. Tokyo

⑧ Innovative Design and Production Processes for the Manufacturing Industry in the Near Future
Inst. of Industrial Science, U. Tokyo

Basic science

②Personalized and Preventive Medicine
Inst. Medical Science, U. Tokyo

④Environmental Predictions with Observational Big Data
Center for Earth Info., JAMSTEC

⑤High-Efficiency Energy Creation, Conversion/Storage and Use
Inst. Molecular Science, NINS

⑦New Functional Devices and High-Performance
Inst. For Solid State Phys., U. Tokyo

⑨Fundamental Laws and Evolution of the Universe
Cent. for Comp. Science, U. Tsukuba

3

Instruction Set Architecture

- **ARM V8 HPC Extension**

  - Fujitsu is a lead partner of ARM HPC extension development

  - Detailed features were announced at Hot Chips 28 - 2016
    http://www.hotchips.org/program/   SVE (Scalable Vector Extension)
    Mon 8/22 Day1 9:45AM GPUs & HPCs
    ARMv8-A Next Generation Vector Architecture for
    HPC

- **Fujitsu's inheritances**

  - FMA

  - Math acceleration primitives

  - Inter core barrier

  - Sector cache

  - Hardware prefetch assist

Post-K : Fujitsu HPC CPU to Support ARM v8   **ARM** **FUJITSU**

Post-K fully utilizes Fujitsu's proven supercomputer microarchitecture

Fujitsu, as a "lead partner" of ARM HPC extension development, is
working to realize an ARM Powered® supercomputer w/ high
application performance

ARM v8 brings out the real strength of Fujitsu's microarchitecture

HPC apps acceleration feature	Post-K	FX100	FX10	K computer
FMA : Floating Multiply and Add	✔	✔	✔	✔
Math. acceleration primitives*	✔Enhanced	✔Enhanced	✔	✔
Inter core barrier	✔	✔	✔	✔
Sector cache	✔Enhanced	✔Enhanced	✔	✔
Hardware prefetch assist	✔Enhanced	✔Enhanced	✔	✔
Tofu interconnect	✔Integrated	✔Integrated	✔	✔

* Mathematical acceleration primitives include trigonometric functions, sine & cosines, and exponential function

2

Copyright 2016 FUJITSU LIMITED

FLAGSHIP2020 Project

- ☐ **Missions**
  - Building the Japanese national flagship supercomputer, Post K, and
  - Developing wide range of HPC applications, running on Post K, in order to solve social and science issues in Japan
- ☐ **Budget**
  - 110 Billion JPY (about 0.91 Billion USD in case of 120 JPY/$)
  - including research, development and acquisition, and application development
- ☐ **Hardware and System Software**
  - Post K Computer
    - RIKEN AICS is in charge of development
    - Fujitsu is vendor partnership

CY	2014	2015	2016	2017	2018	2019	2020	2021	2022
	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4

Basic Design | Design and Implementation | Manufacturing, installation, and Tuning | Operation
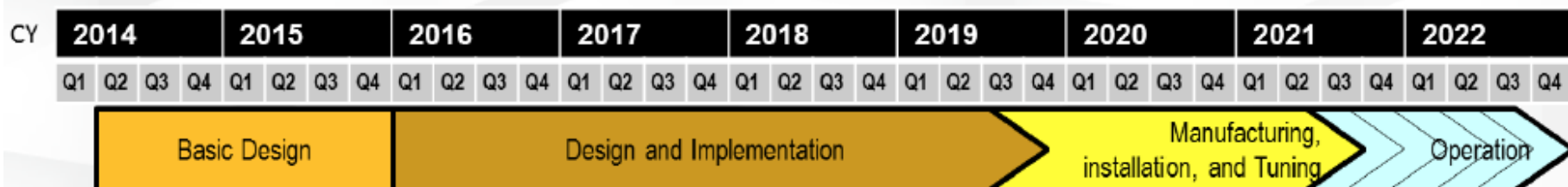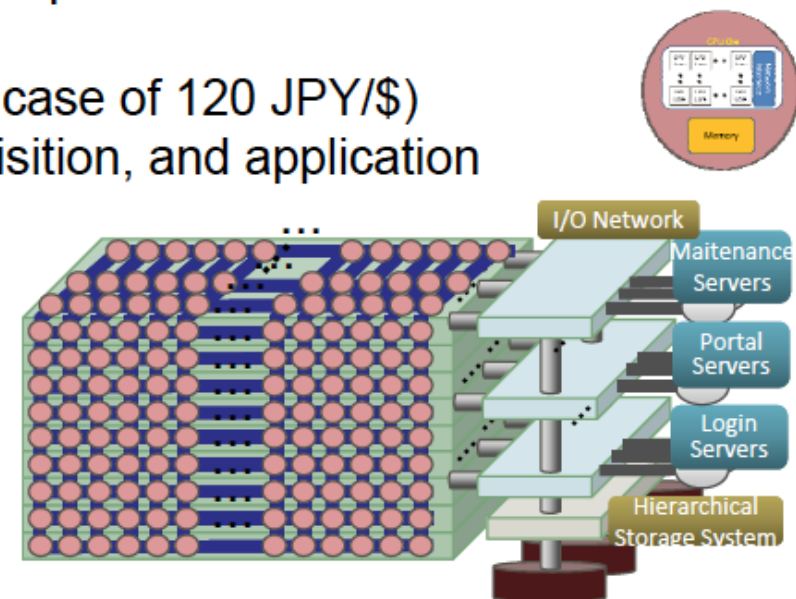
Post K system

- 50 times the K computer for capability computing
- 100 times the K computer for capacity computing
- Power consumption 30-40 MW
- Post K system will go in the same location as K computer
- K computer will have to be dismantled then the Post K system can be installed

CY	2014	2015	2016	2017	2018	2019	2020	2021	2022
	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4

Basic Design — Design and Implementation — Manufacturing, installation, and Tuning — Operation
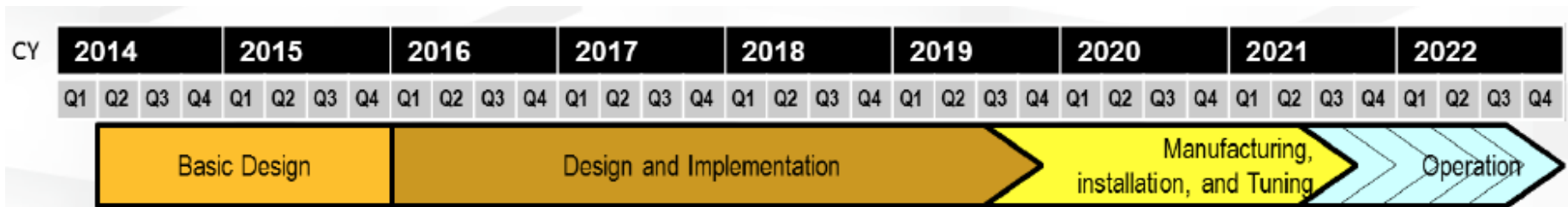
Deployment plan of 9 supercomputing center （Oct. 2015 + latest）

Fiscal Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Hokkaido	HITACHI SR16000/M1（172TF, 22TB) Cloud System BS2000 (44TF, 14TB) Data Science Cloud / Storage HA8000 / WOS7000 (10TF, 1.96PB)				3.2 PF (UCC + CFL/M) 0.96MW / 0.3 PF (Cloud) 0.36MW					30 PF（UCC + CFL-M） 2MW		
Tohoku	NEC SX-9他 (60TF)	SX-ACE(707TF,160TB, 655TB/s) LX406e(31TF), Storage(4PB), 3D Vis, 2MW					~30PF, ~30PB/s Mem BW (CFL-D/CFL-M) ~3MW					
Tsukuba	HA-PACS (1166 TF) COMA (PACS-IX) (1001 TF)	PACS-X 10PF (TPF) 2MW										
Tokyo	Fujitsu FX10 (1PFlops, 150TiB, 408 TB/s), Hitachi SR16K/M1 (54.9 TF, 10.9 TiB, 28.7 TB/s)		Oakforest-PACS 25 PF (UCC + TPF) 4.2 MW / Reedbush 1.8~1.9 PF 0.7 MW				50+ PF (FAC) 3.5MW			100+ PF 4.5MW (UCC + TPF)		200+ PF (FAC) 6.5MW
Tokyo Tech.	TSUBAME 2.5 (5.7 PF, 110+ TB, 1160 TB/s), 1.4MW	TSUBAME 2.5 (3~4 PF, extended)		TSUBAME 3.0 (20 PF, 4~6PB/s) 2.0MW (3.5, 40PF at 2018 if upgradable)						TSUBAME 4.0 (100+ PF, >10PB/s, ~2.0MW)		
Nagoya	FX10(90TF) CX400(470TF) Fujitsu FX100 (2.9PF, 81 TiB) Fujitsu CX400 (774TF, 71TiB) SGI UV2000 (24TF, 20TiB) 2MW in total				50+ PF (FAC/UCC + CFL-M) up to 4MW						100+ PF (FAC/UCC+CFL-M) up to 4MW	
Kyoto	Cray: XE6 + GB8K + XC30 (983TF) Cray XC30 (584TF)	7-8 PF(FAC/TPF + UCC) 1.5 MW					50-100+ PF (FAC/TPF + UCC) 1.8-2.4 MW					
Osaka	NEC SX-ACE NEC Express5800 (423TF) (22.4TF)				3.2PB/s, 5-10Pflop/s, 1.0-1.5MW (CFL-M) / 0.7-1PF (UCC)						25.6 PB/s, 50-100Pflop/s,1.5-2.0MW	
Kyushu	HA8000 (712TF, 242 TB) SR16000 (8.2TF, 6TB) FX10 (272.4TF, 36 TB) CX400 (966.2 TF, 183TB) 2.0MW	15-20 PF (UCC/TPF) 2.6MW FX10 (90.8TFLOPS)					100-150 PF (FAC/TPF + UCC/TPF) 3MW					

Power consumption indicates maximum of power supply (includes cooling facility)

Center for Computational Sciences, Univ. of Tsukuba

Specification of Oakforest-PACS

Total peak performance		25 PFLOPS
Total number of compute nodes		8,208
Compute node	Product	Fujitsu Next-generation PRIMERGY server for HPC (under development)
	Processor	Intel® Xeon Phi™（Knights Landing）Xeon Phi 7250 (1.4GHz TDP) with 68 cores
	Memory　High BW	16 GB, > 400 GB/sec (MCDRAM, effective rate)
	Low BW	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)
Inter-connect	Product	Intel® Omni-Path Architecture
	Link speed	100 Gbps
	Topology	Fat-tree with full-bisection bandwidth
Login node	Product	Fujitsu PRIMERGY RX2530 M2 server
	# of servers	20
	Processor	Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)
	Memory	256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)

Wuxi Workshop 2016, Wuxi China
2016/09/17

Center for Computational Sciences, Univ. of Tsukuba

Schedule

- 2013/7 RFI
- 2015/1 RFC
- 2016/1 RFP
- 2016/3/30 Proposal deadline
- 2016/4/20 Bid opening
- 2016/10/1 1st step system operation (~410 nodes)
- 2016/12/1 2nd step, full system operation
- 2017/4 National open use starts including HPCI
- 2022/3 System shutdown (planned)

Wuxi Workshop 2016, Wuxi China

2016/09/17

Center for Computational Sciences, Univ. of Tsukuba

Partnership for Advanced Computing in Europe

PRACE is an international not-for-profit association under Belgian law, with its seat in Brussels.

PRACE counts 25 members and 2 observers.

The **PRACE** Hosting Members are France, Germany, Italy and Spain.

PRACE is governed by the **PRACE** Council in which each member has a seat. The daily management of the association is delegated to the Board of Directors.

PRACE is funded by its members as well as through a series of implementation projects supported by the European Commission.
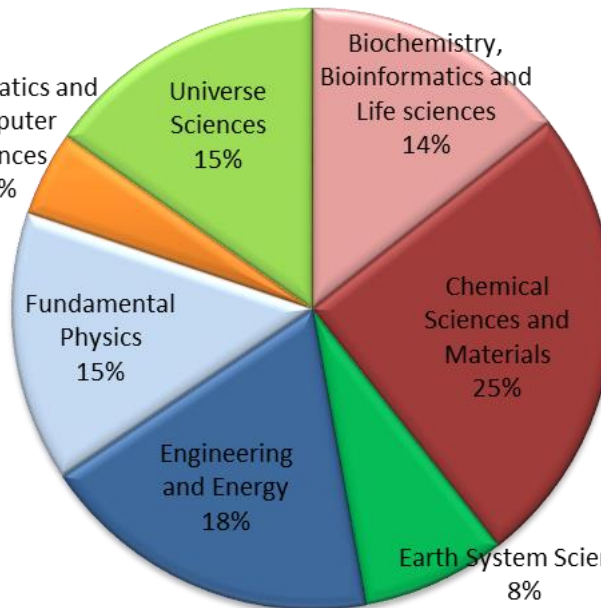
4 Hosting Members offering core hours on 6 world-class machines

JUQUEEN: IBM BlueGene/Q GAUSS/FZJ Jülich, Germany

SuperMUC: IBM GAUSS/LRZ Garching, Germany

Hazel Hen: Cray GAUSS/HLRS, Stuttgart, Germany

MareNostrum: IBM BSC, Barcelona, Spain

CURIE: Bull Bullx GENCI/CEA Bruyères-le-Châtel, France

MARCONI: Lenovo System ,CINECA , Bologna, Italy

Pie chart:
- Biochemistry, Bioinformatics and Life sciences 14%
- Chemical Sciences and Materials 25%
- Earth System Sciences 8%
- Engineering and Energy 18%
- Fundamental Physics 15%
- Mathematics and Computer Sciences 5%
- Universe Sciences 15%

465 scientific projects **enabled**

12.2 thousand million core hours awarded since 2010 with peer review, main criterion is **scientific excellence**. **Open R&D** access for **industrial users** with **>50 companies** supported

>7 350 people trained by **6 PRACE Advanced Training Centers** and others events

21 Pflop/s of peak performance on **6 world-class systems**

530 M€ of funding for **2010-2015**, access **free at the point of usage**

25 members, including **4 Hosting Members** (France, Germany, Italy, Spain)

Criterion: Scientific Excellence

Access through PRACE Peer Review

Free-of-charge required to publish results at the end of the award period
Open to **international** projects

Project Access (*every 6 months*) **award period 1 to 3 years**
Individual researchers and groups
No restriction on nationality for both researcher and centre
Required to demonstrate technical feasibility of project

Preparatory Access (*cut-off date every 3 months*)
Optional support from PRACE experts
Prepare proposals for Project Access

Current rating

The 24th edition of 03.29.2016

N	A place	Max CPU / core	Architecture (Processor type / network)	Performance (Tflop / s) Linpack / Peak	developer
1	Moscow Moscow State University named after MV Lomonosov 2014	1280/37120	units: 1280 (Xeon E5-2697v3 [Acc: Tesla K40M] 2.6 64 GB RAM GHz) Network: FDR Infiniband / FDR Infiniband / Gigabit Ethernet	1,849.00 2,575.87	T-Platforms
2	Moscow Moscow State University named after MV Lomonosov 2012	12422/82468	units: 4160 (2xXeon 5570 2.93 GHz 12 GB RAM) units: 777 (2xXeon E5630 [Acc: 2xTesla X2070] 2.53 GHz 12 GB RAM) units: 640 (2xXeon 5670 2.93 GHz 24 GB RAM) units: 288 (2xXeon E5630 [Acc: 2xTesla X2070] 2.53 GHz 24 GB RAM) units: 260 (2xXeon 5570 2.93 GHz 24 GB RAM) units: 40 (2xXeon 5670 2.93 GHz 48 GB RAM) units: 30 (2xPowerXCell 8i 3.2 GHz 16 GB RAM) Bathrooms: 4 (4xXeon E7650 2.26 GHz 512 GB RAM) network: Infiniband QDR / Gigabit Ethernet / Gigabit Ethernet	901.90 1,700.21	T-Platforms
3	St. Petersburg Supercomputer center of St. Petersburg Polytechnic University " 2014	1424/19936	units: 612 (2xXeon E5-2697v3 2.6 64 GB RAM GHz) units: 56 (2xXeon E5-2697v3 [Acc: 2x NVIDIA K40] 2.6 64 GB RAM GHz) units: 28 (2xXeon E5-2697v3 2.6 GHz 128 GB RAM) Bathrooms: 8 (2xXeon E5-2697v3 [Acc: NVIDIA K1] 2.6 GHz 128 GB RAM) Bathrooms: 8 (2xXeon E5-2697v3 [Acc: NVIDIA K2] 2.6 GHz 128 GB RAM) Network: FDR Infiniband / Gigabit Ethernet / Gigabit Ethernet	658.11 829.34	RSK Group companies

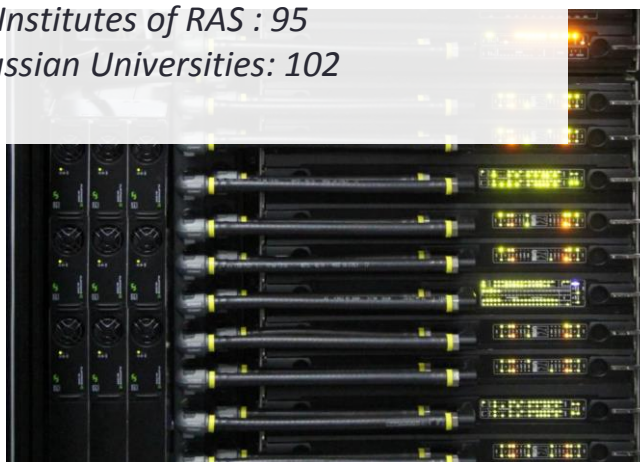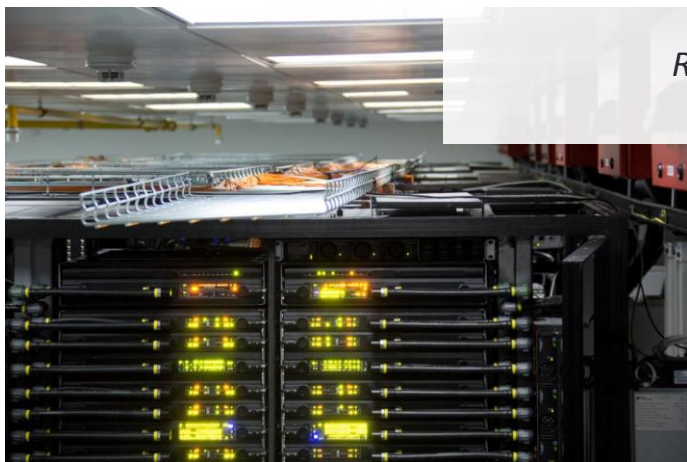# MSU Supercomputer "Lomonosov-2"

MSU Supercomputing Center today:
Users: 2955
Projects: 880

MSU Faculties / Institutes : 21
Institutes of RAS : 95
Russian Universities: 102

1 rack = 256 nodes: Intel (14c) + NVIDIA = 515 Tflop/s
"Lomonosov-2" Supercomputer = 2.9 Pflop/s

Thanks

- **David Kahaner, ATIP**
- **Taisuki Boku, U of Tsukuba**
- **Yutaka Isakawa, RIKEN**
- **Dieter Kranzlmueller, LRZ**
- **Vladimir Voevodin, Moscow State U**
- **Stephane Requena, GENCI**
- **Mateo Valero, Barcelona**
- **The people at the Wuxi Computing Center**